Applying Queueing Theory to Real World Applications

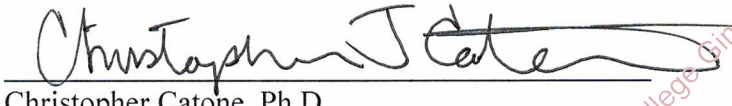
Sophie Bass

Candidate for the degree

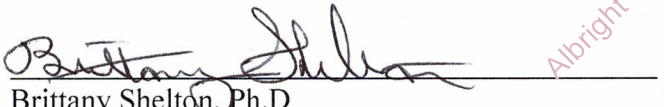
Bachelor of Sciences

Submitted in partial fulfilment of the requirements for

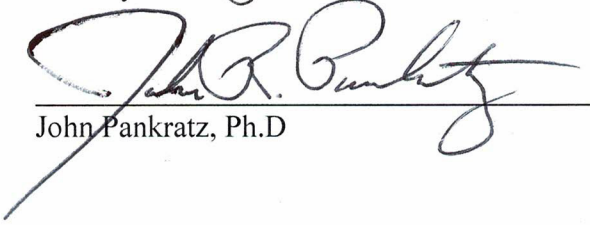
College Honors



Christopher Catone, Ph.D



Brittany Shelton, Ph.D



John Pankratz, Ph.D

F. Wilbur Gingrich Library
Special Collections Department
Albright College

Release of Senior Thesis

I hereby grant to the Special Collections Department of the F. Wilbur Gingrich Library at Albright College the nonexclusive right and privilege to reproduce, disseminate, or otherwise preserve the Senior Honors Thesis described below in any noncommercial manner that furthers the educational, research or public service purposes of Albright College. Copyright privileges remain with me, the author.

Title: Applying Queueing Theory to Real World Applications

Signature of Author: Sophie Bass Date: 4/17/19

Printed Name of Author: Sophie Bass

Street Address: 15 North Wickom Drive

City, State, Zip Code: Westfield, NJ 07090

Applying Queueing Theory to Real World Applications

Senior Honors Thesis, Albright College

Sophie Bass

Table of Contents

I. Abstract

This thesis is about the study of Queueing Theory and applying this theory to real-world applications. In this project, we addressed what queueing theory is, the basic structure of queueing models and different types of real queueing systems. We also focused on the two important distributions: Poisson and Exponential and displayed queueing systems through single and multiple server cases. We conclude this thesis with real-world examples of Queueing theory.

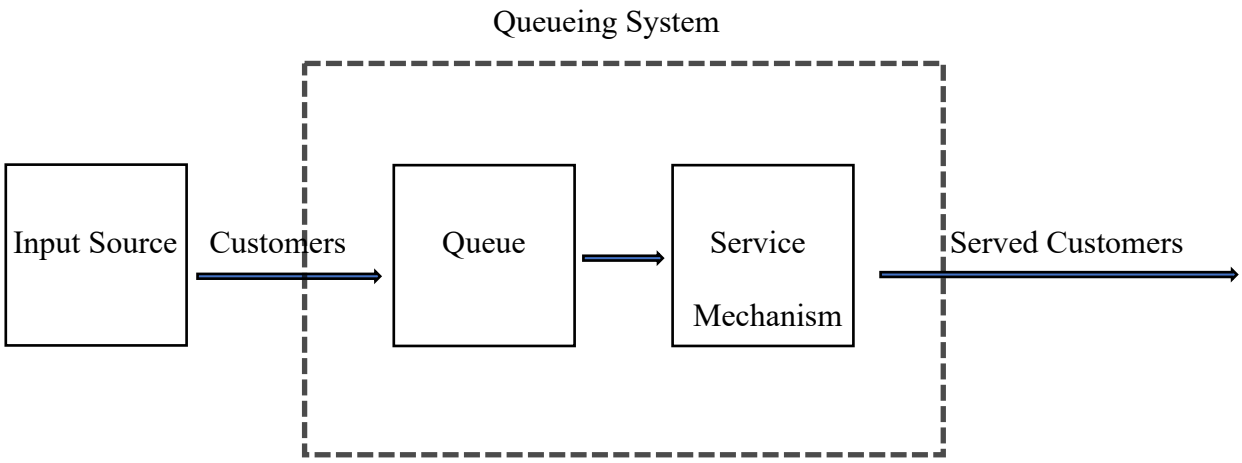
II. Introduction

During the last century, times have changed the way people receive services in all different aspects, including the grocery store, gas stations, emergency rooms, toll booths and much more. This mathematical study is called queueing theory and is the study of queues or waiting lines. Every day one will encounter queueing theory without even knowing it. The ultimate goal of queueing theory is to achieve a balance between the cost associated with waiting for the service and the cost of the service itself. When people encounter lines, they ask themselves simple questions that might help them determine which line they should stand in. Some of these questions are "how long will the line be, how long will the wait be, how busy will the server who is servicing the line be, and how much capacity is needed to meet an expected level of demand" (Schwartz)? However, we also can look at this from the server's point of view and they can wonder "how likely am I to lose business due to too-long waits and not enough capacity and how much more demand can we satisfy without creating an unacceptably long wait" (Schwartz)?

There are many cons to excessive waiting and long lines, which include the loss of customers and the social cost for customers. However, "providing too much service would involve excessive costs. On the other hand, not providing enough service capacity would cause the waiting line to become excessively long at times" (Hiller 379). I think it is important to understand queueing theory from a mathematical standpoint and this thesis is here to contribute knowledgeable information required to help companies figure out the average time someone waits on a line through different characteristics.

III. Basic Structure of Queueing Models

The basic process of a queueing system consists of the input source, the queue, service discipline and service mechanism. Below we explain each step in the queueing system. The diagram illustrates how the steps are interrelated.



Input Source

An input source can be described as the calling population or a population of individuals which are entering the queue. The size of an input source is the total number of potential customers who may need assistance with service from time to time. The input source can be either finite or infinite depending on the characteristics of the queue. The infinite case is generally easier to study than the finite case. There are two main processes and distributions used in queueing theory: the Poisson distribution and the exponential distribution. "The statistical pattern by which customers are generated over time is according to a Poisson process, which the number of customers generated until any specific time has a Poisson distribution" (Hiller 381). While the exponential distribution showcases an equivalent assumption that displays the probability distribution of the time between consecutive arrivals also known as the interarrival time.

Queue

Queues can be infinite, or finite depending on the system and can be described by the number of potential customers that it contains. An example of a finite queue is limited seating arrangements in a restaurant. When the queue is infinite, the modeling process is simpler. An example of an infinite queue is people checking out at a supermarket.

Service Discipline

Service discipline is the order in which customers in the queue are selected for service. There are a lot of different queue disciplines that are possible. The one that is normally used is first-come-first-served which can be described as the process in the order which you arrive in the queue is the order that you will be helped. Most companies assume that first-come-first-served is the most fair and equal for the customer. Some other queueing disciplines include priority procedure and random. Priority procedure can be described as every item has a priority associated with it and customers that have a high priority will be serviced before the customers with the lower priority. The last way customers can be selected in a queue is randomly which is when the order in which customers are serviced is random.

Service Mechanism

The next step in the queueing system is service mechanism. The service mechanism is the way that customers receive service once they are selected from the front of a queue. Service mechanisms consists of a number of facilities to help potential customers and they may also be called a server. Customers may receive service from multiple facilities, if there is more than one

facility and more than one server. The waiting time in line before being helped is known as the holding time. A service mechanism can also be called a server and more queueing models assume that there is a single server.

Elementary Queueing Process

The most basic queueing theory model consists of one line in which customers wait for a single service facility, with one or more servers. Once a customer gets up to the front of the line, they will be called by a server when the person in front of them is finished. Therefore, if the line is long at the beginning one should know that this line will break into however many servers there are in the queue. A server can be one-person, multiple people or sometimes it may not be a person. A server can also be a machine or a piece of equipment. Sometimes the customers who are waiting in the queues are not even people. A customer could be an item waiting for a certain operation by a server, which could be a machine. Another example of servers and customers is cars waiting at a toll booth. “When analyzing queueing theory, the data includes the average number waiting to be served and the average waiting time because it is irrelevant whether or not the customers wait together in individually versus a group ” (Hiller 383).

IV. Types of Real Queueing Systems

There are many different important classes of queueing systems that one encounters on a day to day basis. The first important class of queueing systems is commercial service systems, which is when customers will use commercial organizations for their service. This type of queueing system is mostly person-to-person service at a set location. Some examples include a grocery store, hair salon, bank teller or a cafeteria line. Also, if someone comes to fix something in ones house, the server, which is the repairman, travels to the customers house to complete the repair. Some examples that are not considered commercial service systems, include gas stations because the customers are the cars.

The next important class of queueing systems is transportation service systems. For this case, the vehicles, which include cars, airplanes, and ships are the customers and the servers are the toll booths and runways. One example is a parking lot where the customers are the cars and the servers are the parking spaces. However, this is not technically a queue because the cars must go elsewhere to park if the parking lot is full.

In the most recent years, queueing theory has started to be applied mostly to business industrial service systems. Some examples of this type of queueing system include inspection stations, maintenance systems and materials handling systems. For inspection systems, the inspectors are the servers and the items they inspect are the customers. There are many more examples that are similar to the inspection system, which describe who the customer and servers are in their respective situations.

The last type of queueing system that is now growing is social service systems. Examples of social service systems include judicial systems, legislative systems, and health care systems. For

judicial systems the judges are the servers and the cases waiting to be judged are the customers. Also, hospitals and incoming hospital vehicles can be viewed as queueing systems where ambulances and hospital beds are the servers and families waiting for service are the customers.

Although, these are four cases that we mention in detail, there are still more types of queueing systems that are not listed. Queueing theory was first studied with telephone engineering and still is changing and expanding to this day.

V. Two Important Distributions

Queueing theory utilizes two important distributions; Poisson and Exponential. The characteristics of queueing systems are determined by two major properties, the probability distribution of service times and the probability distribution of interarrival times.

The Role of Poisson Distribution

The Poisson distribution was named after S.D. Poisson, a French mathematician and can be defined as "the relative frequency distribution of the number of rare events that occur randomly in a specified unit of space, distance, or time" (Mendenhall 157). The Probability Mass Function for a Poisson distribution and is given by

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

where,

μ *is the mean value of* x

and

x *is a nonnegative integer that counts the number of rare events observed* $(x = 0,1,2,3 \ldots)$.

The Role of Exponential Distribution

Now, suppose a random variable T represents either interarrival or service times. We will recall that T can be said to be exponential with parameter $\propto$ when the probability density function is

$$f_T(t) = \begin{cases} \propto e^{-\propto t} & for\ t \geq 0 \\ 0 & for\ t > 0 \end{cases}$$

Thus, the cumulative probabilities are

$$P(T \leq t) = 1 - e^{-\propto t} \quad \text{for } t \geq 0$$

$$P(T \geq t) = e^{-\propto t} \quad \text{for } t \geq 0$$

and the expected value and variance of T are

$$E(T) = \frac{1}{\propto} \text{ and } var(T) = \frac{1}{\propto^2}.$$

There are five key properties of the exponential distribution.

Property 1: $f_T(t)$ is a strictly decreasing function of t ($t \geq 0$)

Property 1 states $P(0 \leq T \leq \Delta t) > P(t \leq T \leq t + \Delta t)$

for any strictly positive values of Δt and t. Thus, it is relatively likely that T will take on a small value near zero and

$$P\left(0 \leq T \leq \frac{1}{2}\frac{1}{\propto}\right) = .393$$

whereas

$$P\left(\frac{1}{2}\frac{1}{\propto} \leq T \leq \frac{3}{2}\frac{1}{\propto}\right) = .383$$

and thus, the value T is more likely small and is decreasing.

Property 2: Lack of Memory

This property can be stated mathematically as

$$P(T > t + \Delta t | T > \Delta t) = P(T > t)$$

for any positive quantities t and Δt. "In other words, this can be described as the probability distribution of the remaining time until the incident occurs always is the same, regardless of how much time has passed. In effect, the process "forgets" its history" (Hiller 388). Thus,

$$P(T > t + \Delta t | T > \Delta t) = \frac{P(T > \Delta t, T > t + \Delta t)}{P(T > \Delta t)}$$

$$= \frac{P(T > t + \Delta t)}{P(T > \Delta t)}$$

$$= \frac{e^{-\propto(t+\Delta t)}}{e^{-\propto \Delta t}}$$

$$= e^{-\propto t}$$

Property 3: The minimum of several independent exponential random variables has an exponential distribution

This property can be stated mathematically; let $T_1, T_2, \dots, T_n$ be independent exponential random variables with parameters $\propto_1, \propto_2, \dots, \propto_n$ respectively. Also, let U be the random variable that takes on the value equal to the minimum of the values actually taken on by $T_1, T_2, \dots, T_n$; that is

$$U = \min (T_1, T_2, \dots, T_n) .$$

Thus, if T_i represents the time until a particular kind of incident will occur, then U represents the time until the first of the *n* different incidents will occur. Now to show that the random variables has an exponential distribution for any $t \geq 0$,

$$P(U > t) = P(T_1 > t, T_2 > t, \dots, T_n > \mathrm{t})$$

$$= P(T_1 > t)P(T_2 > t) \dots P(T_n > t)$$

$$= e^{-\propto_1 t} e^{-\propto_2 t} \dots e^{-\propto_n t}$$

$$= exp\left(-\sum_{i=1}^{n} \propto_i t\right)$$

So that U has an exponential distribution with parameter

$$\propto = \sum_{i=1}^{n} \propto_i$$

Property 4: Relationship to the Poisson distribution

This property has to do with the resulting solution about the probability distribution of the number of times this kind of incident occurs over a specified length of time. Thus, let X(t) be the number of occurrences by time t ($t > 0$), with time 0 displays when the time count should begin. The solution is

$$P(X(t) = n) = \frac{(\propto t)^n e^{-\propto t}}{n!}, for\ n = 0, 1, 2\ldots$$

that is, $X(t)$ has a Poisson distribution with parameter $\propto t$.

Property 5: For all positive values of t, P($T \leq t + \Delta t | T > t$) $\approx \propto \Delta t$ for small Δt.

Since the series expansion of e^x for an exponent x is $e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}$

it follows that $P(T \leq t + \Delta t | T > t) = 1 - 1 + \propto \Delta t - \sum_{n=2}^{\infty} \frac{(-\propto \Delta t)^n}{n!}$

$\approx \propto \Delta t$, for small Δt.

VI. Birth and Death Process

By birth and death processes, we are modeling the arrival of customers entering the system and customers leaving the system. Customers entering the system must exit. Hence at any state the rate in equals rate out. By state, we mean the number of customers in the system. The equations which describe these rates are called balance equations.

State	Rate In=Rate Out
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$
2	$\lambda_1 P_1 + \mu_2 P_3 = (\lambda_2 + \mu_2) P_2$
…	…
n-1	$\lambda_{n-2} P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1}) P_{n-1}$
n	$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$
…	…

P_n= probability that exactly n customers are in the system
μ_n= mean service rate for n customers overall in the system
λ_n= mean arrival rate of n customers overall in the system

Solving the balance equations for P_n yields the following:

State

0: $P_1 = \frac{\lambda_0}{\mu_1} P_0$

1: $P_2 = \frac{\lambda_1}{\mu_2} P_1 + \frac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$

2: $P_3 = \frac{\lambda_2}{\mu_3} P_2 + \frac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) = \frac{\lambda_2}{\mu_3} P_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0$

…

n-1: $P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} + \frac{1}{\mu_n}(\mu_{n-1} P_{n-1} - \lambda_{n-2} P_{n-2}) = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = \frac{\lambda_{n-1} \lambda_{n-2} \dots \lambda_0}{\mu_n \mu_{n-1} \dots \mu_1} P_0$

n: $P_{n+1} = \frac{\lambda_n}{\mu_n + 1} P_n + \frac{1}{\mu_{n+1}}(\mu_n P_n - \lambda_{n-1} P_{n-1}) = \frac{\lambda_n}{\mu_{n+1}} P_n = \frac{\lambda_n \lambda_{n-1} \dots \lambda_0}{\mu_{n+1} \mu_n \dots \mu_1} P_0$

…

And to simplify, let

$$C_n = \frac{\lambda_{n-1} \lambda_{n-2} \dots \lambda_0}{\mu_n \mu_{n-1} \dots \mu_1} \; for \; n = 1,2, \dots$$

and hence $P_n = C_n P_0$. (Where P_n is probability that exactly n customers are in queuing system at time t, given number at time 0.)

VII. Single Server Case

Basic Model with Infinite Queue

It is common for the mean service rate per busy server and the mean arrival rate for the queueing systems to be essentially constant. In this case, one can use the basic model to describe this queueing system. Let the system have just a single server (s=1) and assume the parameters for the birth and death processes are

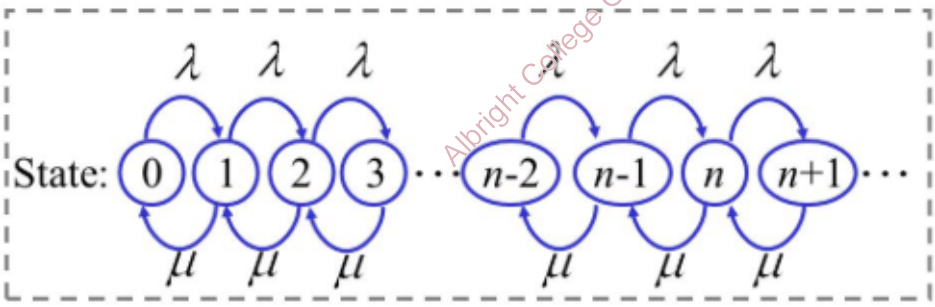
$$\lambda_n = \lambda \text{ for } n = 0,1,2,\ldots$$

where λ is the mean arrival rate of new customers when n customers are in a system and

$$\mu_n = \mu \text{ for } n = 1,2,\ldots$$

where μ is the mean service rate for overall system when n customers are in a system.

This assumes that the rate of the arrival and rate of service are not dependent on the number of customers waiting in the queue. We can visualize the system using a rate diagram, which is a chart that summarizes the given information for a queueing theory problem. The arrows in the diagram show the possible transitions for each state and the label for each arrow gives the mean rate for that transition when the system is in the state at the base of the arrow.



For a single server (s=1), the C_n factors for the birth and death process reduce to

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n, \text{ for n=1,2,…}$$

Therefore,

$$P_n = \rho^n P_0 \text{ for n=1,2,…,}$$

where P_n is the probability that exactly *n* customers are in queueing system and

$$P_0 = \frac{1}{1+\sum_{n=1}^{\infty} \rho^n}$$

$$=(\sum_{n=0}^{\infty} \rho^n)^{-1}$$

$$=\left(\frac{1}{1-\rho}\right)^{-1}$$

$$=1-\rho.$$

Thus,

$$P_n = (1-\rho)\,\rho^n, \text{ for n=0,1,2,…}$$

Consequently, now we will define the formula for L, which is the expected number of customers in a queueing system.

$$L = \sum_{n=0}^{\infty} nP_n$$

$$= \sum_{n=0}^{\infty} n(1-\rho)\,\rho^n$$

$$= (1-\rho)\,\rho \sum_{n=0}^{\infty} n\rho^{n-1}$$

$$= (1-\rho)\,\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}(\rho^n)$$

$$= (1-\rho)\,\rho \frac{d}{d\rho} \sum_{n=0}^{\infty} \rho^n$$

$$= (1-\rho)\,\rho \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

$$= \frac{\rho}{1-\rho}$$

$$= \frac{\lambda}{\mu-\lambda}$$

Similarly, we will define L_q, which is the expected queue length. The expected queue length is the length of the queue when all the customers are in line.

We have,

$$L_q = \sum_{n=1}^{\infty}(n-1)P_n$$

$$= \sum_{n=1}^{\infty} nP_n - \sum_{n=1}^{\infty} P_n$$

$$= L - (1 - P_0)$$

$$= L - 1 + P_0$$

$$= L - 1 + 1 - \rho$$

$$= \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu}$$

$$= \frac{\lambda^2}{\mu(\mu - \lambda)}$$

Since the probability that the random arrival will find n customers in the system is P_n, it follows that

$$\text{P}(W > \text{t}) = \sum_{n=0}^{\infty} P_n P(S_{n+1} > t)$$

[where S_{n+1} is known to have a gamma distribution with the cumulative distribution function].

This will reduce to

$$\text{P}(W > \text{t}) = e^{-\mu(1-\rho)t} \; for \; t \geq 0$$

where W is the expected waiting time in the system including service time when the service is first-come-first-served.

Also,

$$P(W_q > t) = \rho e^{-\mu(1-\rho)t}, for \; t \geq 0$$

$$W_q = E(W_q) = \frac{\lambda}{\mu(\mu-\lambda)}$$

where $\mathcal{W}_q$ is the expected waiting time in the system excluding service time for any random variable.

Basic Model with a Finite Queue

If the queue is known to be finite, the number of customers in the system is not allowed to exceed a specific number and we denote this number by M. This case is similar to the infinite input source case. We only need to change the parameters which describe the mean arrival rate to

$$\lambda_n = \begin{cases} \lambda, for\ n = 0,1,2, \dots M-1 \\ 0, for\ n \geq M \end{cases}$$

Thus when s=1, $C_n = \begin{cases} \left(\frac{\lambda}{\mu}\right)^n = \rho^n, for\ n = 1,2, \dots, M \\ 0 \qquad\qquad\qquad for\ n > M \end{cases}$

Therefore, $P_0 = \frac{1}{\sum_{n=0}^{M}(\lambda/\mu)^n}$

$= 1/\left[\frac{1-(\lambda/\mu)^{M+1}}{1-(\lambda/\mu)}\right]$

$= \frac{1-\rho}{1-\rho^{M+1}}$

where P_0 is the probability that exactly 0 customers are in the queueing system.

Thus,

$$P_n = \left(\frac{1-\rho}{1-\rho^{M+1}}\right)\rho^n \quad \text{for n=0,1,2,…M.}$$

Now, we will solve for L to find the expected number of customers in the queueing system.

$$L = \sum_{n=0}^{M} nP_n$$

$$= \frac{1-\rho}{1-\rho^{M+1}}\rho \sum_{n=0}^{M} \frac{d}{d\rho}(\rho^n)$$

$$= \frac{1-\rho}{1-\rho^{M+1}}\rho \frac{d}{d\rho}\sum_{n=0}^{M} \rho^n$$

$$= \left(\frac{1-\rho}{1-\rho^{M+1}}\right)\rho \frac{d}{d\rho}\left(\frac{1-\rho^{M+1}}{1-\rho}\right)$$

$$= \rho \frac{-(M+1)\rho^M + M\rho^{M+1} + 1}{(1-\rho^{M+1})(1-\rho)}$$

$$= \frac{\rho}{1-\rho} - \frac{(M+1)\rho^{M+1}}{1-\rho^{M+1}}.$$

VIII. Multiple Server Case

Let the system has multiple servers (k > 1). As in the previous case assume the parameters for the birth and death processes are

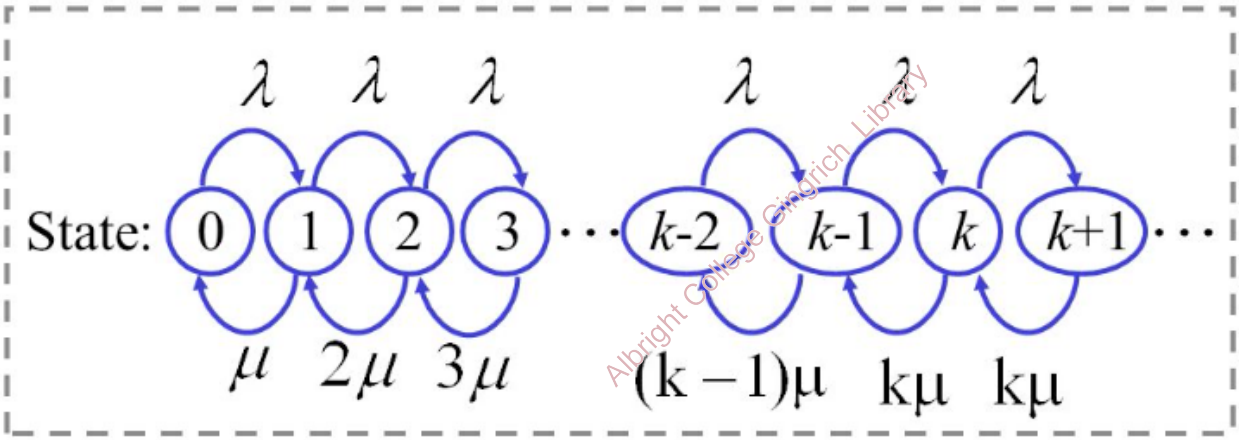
$$\lambda_n = \lambda (n=0,1,2,\ldots)$$

where λ is the mean arrival rate of new customers when n customers are in the system and

$$\mu_n = \begin{cases} n\mu \; for \; n = \; 1,2,\ldots,k \\ k\mu \; for \; n = k, k+1, \ldots \end{cases}$$

where μ is the mean service rate for overall system when n customers are in a system.

For the single server case, I mentioned that one could use a rate diagram to summarize the information for a queueing theory problem and this is also true for multiple server cases. The only difference is the change of mean service rates. In this case, our rate diagram takes the following form.



For multiple servers when s>1, M is the maximum number of servers that could be used and thus we will assume s≤M. Thus, C_n becomes

$$C_n = \begin{cases} \dfrac{\left(\frac{\lambda}{\mu}\right)^n}{n!} \; for\; n = 1,2, \dots s \\ \dfrac{\left(\frac{\lambda}{\mu}\right)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^{n-s} = \dfrac{\left(\frac{\lambda}{\mu}\right)^n}{s!\, s^{n-s}} \; for\; n = s, s+1, \dots \end{cases}$$

Consequently, if $\lambda < s\mu$, then

$$P_0 = 1/\left[\sum_{n=0}^{s-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} \sum_{n=s}^{\infty} \left(\frac{\lambda}{s\mu}\right)^{n-s}\right]$$

$$= 1/\left[\sum_{n=0}^{s-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} \frac{1}{1-\left(\frac{\lambda}{s\mu}\right)}\right]$$

and

$$P_n = \begin{cases} \dfrac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 & for\; n = 1,2, \dots s \\ \dfrac{\left(\frac{\lambda}{\mu}\right)^n}{s!\, s^{n-s}} P_0 & for\; n = s, s+1, \dots M \\ 0 & for\; n > M. \end{cases}$$

We then have,

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2}$$

$$W_q = \frac{L_q}{\lambda}$$

$$W = W_q + \frac{1}{\mu}$$

$$L = \lambda\left(W_q + \frac{1}{\mu}\right) = L_q + \frac{\lambda}{\mu}.$$

IX. Examples of Queueing Theory

Now we are going to consider some real-life examples of queueing theory and calculate the relevant variables for each example.

A grocery store on Westfield Ave has a single checkout stand with a full-time cashier manning it. Customers arrive at the stand "randomly" at a mean rate of 30 per hour. When there is only one customer at the stand, he/she is processed by the cashier alone, with an expected service time of 1.5 minutes. However, the stock boy has been given specific instructions that whenever there is more than one customer at the stand he is to help the cashier by boxing the groceries. This reduces the expected time required to process a customer to 1 minute. In both cases the service time distribution is exponential and the arrival rate is Poisson.

For this example, we are going to state the steady-state probability distribution of the number of customer at the checkout stand.

$$P_0 = \frac{1}{1+\sum_{n=1}^{\infty}\frac{\lambda^n}{\mu_1{\mu_2}^{n-1}}} = \frac{1}{1+\frac{\lambda}{\mu_1}\sum_{n=1}^{\infty}\left(\frac{\lambda}{\mu_2}\right)^{n-1}} = \frac{1}{1+\frac{\lambda}{\mu_1}\left(\frac{1}{1-\lambda/\mu_2}\right)} = \frac{1}{1+30/40(1/(1-\frac{30}{60})} = 2/5$$

$$P_n = P_0 * \frac{\lambda_n}{\mu_1{\mu_2}^{n-1}} = \left(\frac{2}{5}\right) * \frac{30^n}{(40)(60)^{n-1}} = (.6)(\tfrac{1}{2})^n$$

Next, we are going to derive L for this system and use this information to determine $L_q, W, and\ W_q$.

$$L = \sum_{n=0}^{\infty} n * P_n = .6 * \sum_{n=1}^{\infty} n * \left(\frac{1}{2}\right)^n$$

$$= (.6)(\tfrac{1}{2})\sum_{n=1}^{\infty} n * \left(\frac{1}{2}\right)^{n-1}$$

$$= (.6)\left(\frac{1}{2}\right)\left(\frac{1}{1-\frac{1}{2}}\right)^2 = \frac{6}{5}$$

$$L_q = L - (1 - P_0) = \frac{6}{5} - (1 - .4) = \frac{3}{5}$$

$$W = \frac{L}{\lambda} = \frac{\frac{6}{5}}{30} = \frac{1}{25}$$

$$W_q = \frac{L_q}{\lambda} = \frac{\frac{3}{5}}{30} = \frac{1}{50}$$

Therefore, we have found that the probability that exactly *n* customers are in the system at a given time is $(.6)(\frac{1}{2})^n$. Using this information, we found the expected number of customers in the system at the grocery store was 6/5, the expected queue length was 3/5, the expected waiting time including service time is 1/25 minutes and the expected waiting time excluding service time is 1/50 minutes.

Next, we will consider a second situation. Suppose that one repairman has been assigned the responsibility of maintaining three washing machines. For each machine the probability distribution of the running time before a breakdown is exponential, with a mean of 9 hours. The repair time also has an exponential distribution with a mean of 2 hours.

First, we are going to calculate the steady-state probability distribution and the expected number of washing machines that are not running.

Given:
$\lambda = \frac{1}{9}\ per\ hour$
$\mu = \frac{1}{2}\ per\ hour$
$\rho = \frac{\lambda}{\mu} = 2/9$

$$C_0 = 1$$

$$C_1 = \frac{3/9}{1/2} = \frac{2}{3}$$

$$C_2 = \frac{3/9}{1/2} * \frac{2/9}{1/2} = \frac{24}{81}$$

$$C_3 = \frac{3/9}{1/2} * \frac{2/9}{1/2} * \frac{1/9}{1/2} = 16/243$$

$P_0 = \frac{1}{1+\sum C_n}$=1+2/3+24/81+16/243=.4929

$P_1 = P_0 * C_1 = .4929 * 2/3 = .329$

$$P_2 = P_0 * C_2 = .4929 * 24/81 = .146$$

$$P_3 = P_0 * C_3 = .4929 * 16/243 = .032$$

Now using the above information from the second situation, it can be assumed that the calling population is infinite, so that the input process is Poisson with a mean arrival rate of three every 9 hours. One should compare the result from part a with that obtained by making the approximation using the corresponding infinite queue model.

Given:
$\lambda = 1/3\ per\ hour$
$\mu = 1/2\ per\ hour$
$\rho = 2/3$

$$P_0 = 1 - \rho = 1 - \frac{2}{3} = 1/3$$
$$P_n = \rho^n * P^0 = 1/3 * \rho^n$$

$$P_1 = \frac{1}{3} * \frac{2}{3} = \frac{2}{9}$$
$$P_2 = \frac{1}{3} * \left(\frac{2}{3}\right)^2 = \frac{4}{27}$$
$$P_3 = \frac{1}{3} * \left(\frac{2}{3}\right)^3 = \frac{8}{81}$$
$$P(n > 3) = 1 - P(n \leq 3) = 1 - \left(\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \frac{8}{81}\right) = .1976$$

Consequently, we have shown that when the breakdown mean is 9 hours and the repair time mean is 2 hours then the probability that there are exactly 0 customers in the system is .4929. 1 customer is .329, 2 customers is .146, and 3 customers is .032. However, when the mean arrival rate changes to three every 9 hours, the probabilities alter a bit. Thus, when there are exactly 0 customers in the system the probability is 1/3, 1 customer is 2/9, 2 customers is 4/27 and three customers is 8/81.

The third situation that we considered is the town of Westfield would like to open a carwash operation on Grove Street and the decision must be made as to how much space to provide to the cars that are waiting. It is estimated that the customers would arrive randomly (Poisson input process) with a mean rate of one every 4 minutes. The time that can be attributed to washing one

car has an exponential distribution with a mean of 3 minutes. I would like you to compare the expected fraction of potential customers that would be lost because of inadequate waiting space if there is one, three, or five spaces provided.

Given:
$\lambda = \frac{1}{4}$
$\mu = 1/3$
$\rho = \frac{\lambda}{\mu} = \frac{1/4}{1/3} = 3/4$
$P_k = \frac{(1-\rho)}{(1-\rho^{k+1})}\rho^k$
K=1,3,5

If there is one space provided then,

$$P_1 = \frac{(1-3/4)}{(1-3/4^2)} * \frac{3}{4} = .429$$

If there is three spaces provided then,

$$P_3 = \frac{(1-3/4)}{(1-3/4^4)} * \left(\frac{3}{4}\right)^3 = .154$$

If there is five spaces provided then,

$$P_5 = \frac{(1-3/4)}{(1-3/4^6)} * \left(\frac{3}{4}\right)^5 = .072$$

Consequently, we have concluded when there are less spaces provided there is a higher number of customers lost due to inadequate waiting. When there is only one space provided the number of potential customers that would be lost is .429 versus when there are five spaces provided the number of potential customers that be lost is .072. Thus, we can see that when there are more spaces, there is less likely chance for potential customers to be lost.

X. Terminology and Notation

Balking = when a customer refuses to enter the system and is lost when the queue is too long

State of the system = number of customers in queueing system

Queue Length = number of customers waiting for service or state of system minus number of customers being served

N(t) = numbers of customers in queuing system at time t ($t \geq 0$)

$P_n(t)$ = probability that exactly n customers are in queuing system at time t, given the number of customers at time 0

s = number of servers (parallel service channels) in queuing system

λ_n = mean arrival rate (expected number of arrivals per unit time) of new customers when n customers are in system

μ_n = mean service rate for overall system (expected number of customers completing service per unit time) when n customers are in system

P_n = probability that exactly n customers are in queueing system

L = expected number of customers in queueing system

L_q = expected queue length

W = expected waiting time in system (includes service time)

W_q = expected waiting time in queue (excludes service time)

XI. Acknowledgements

First, I would like to thank Professor Catone for convincing me to complete this honors thesis in Spring 2018. Without him reaching out to me and helping me choose a topic to research, I would have not gone through with the last step of the Honors Program. When we met in the spring time he offered me 3 different topics that would be interesting to research and I chose queueing theory because I thought it related to our life every day. I also would like to thank my readers Professor Shelton and Professor Pankratz for reading over my drafts and giving me advice on this project.

Lastly to my family, because without them I would not have been able to complete this without the support. There were lots of times where I thought I would never get the section done before each due date, but my mom always reminded me that it would get done at some point. This process was challenging for me because I am in season with lacrosse, but putting time aside each day to work on my thesis and not waiting until one day to do the whole thing was one of the best ways to approach the Honors Thesis Process.

XII. References

Hillier, Frederick Stanton., and Gerald J. Lieberman. *Operations Research*. Holden-Day, 1976.

Mendenhall, William. *Introduction to Probability and Statistics*. Duxbury Pr., 1984.

Schwartz, Baron. "The Essential Guide to Queueing Theory." *The Essential Guide to Queueing Theory*, 8 Sept. 2016, www.percona.com/live/17/sites/default/files/the-essential-guide-to-queueing-theory.pdf.